

## (12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization International Bureau



(43) International Publication Date  
19 August 2004 (19.08.2004)

PCT

(10) International Publication Number  
WO 2004/070627 A1

(51) International Patent Classification<sup>7</sup>: G06F 17/30

(21) International Application Number: PCT/GB2004/000143

(22) International Filing Date: 20 January 2004 (20.01.2004)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data: 0303018.6 10 February 2003 (10.02.2003) GB

(71) Applicant (for all designated States except US): BRITISH TELECOMMUNICATIONS PUBLIC LIMITED COMPANY [GB/GB]; 81 Newgate Street, London, Greater London EC1A 7AJ (GB).

(72) Inventors; and

(75) Inventors/Applicants (for US only): CASE, Simon, James [GB/GB]; 4 Alder Close, Benhall Green, Saxmundham, Suffolk IP17 1QB (GB). FISHER, Michelle, Jayne [GB/GB]; Flat 4, 98 Mount Pleasant Road, Exeter, Devon EX4 7AB (GB).

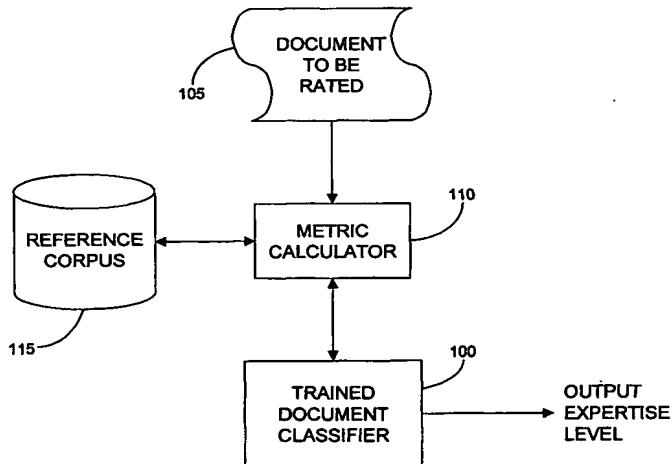
(74) Agent: LLOYD, Barry, George, William; BT Group Legal, Intellectual Property Department, PPC5A, Bt Centre, 81 Newgate Street, London, Greater London EC1A 7AJ (GB).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

[Continued on next page]

(54) Title: DETERMINING A LEVEL OF EXPERTISE OF A TEXT USING CLASSIFICATION AND APPLICATION TO INFORMATION RETRIEVAL



(57) Abstract: An apparatus and method are provided for determining a level of expertise applicable to a particular document and for using this determined level of expertise in an improved information retrieval arrangement. A trainable document classifier is used to identify an applicable level of expertise using a metric indicative of the commonality, as measured with reference to a reference corpus, of terms comprised in a given document, trained using a training set of documents comprising, for each of a plurality of predetermined levels of expertise, a representative sample of documents and their respective metric values. An information retrieval apparatus is arranged to identify documents relevant to a specified category of information and to select from documents so identified those having a level of expertise, determined by the trained document classifier, matching a specified level of expertise for a target user in respect of that category of information.

WO 2004/070627 A1



**Published:**

— *with international search report*

*For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.*